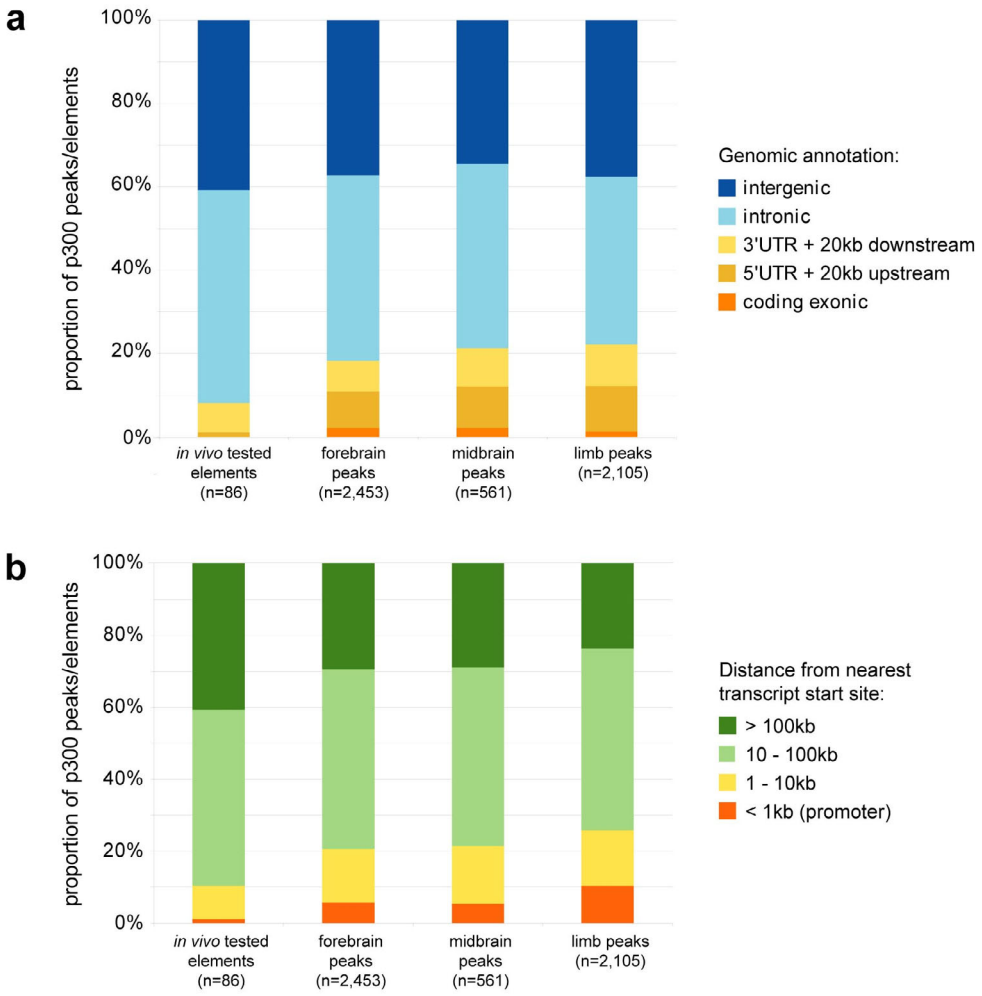
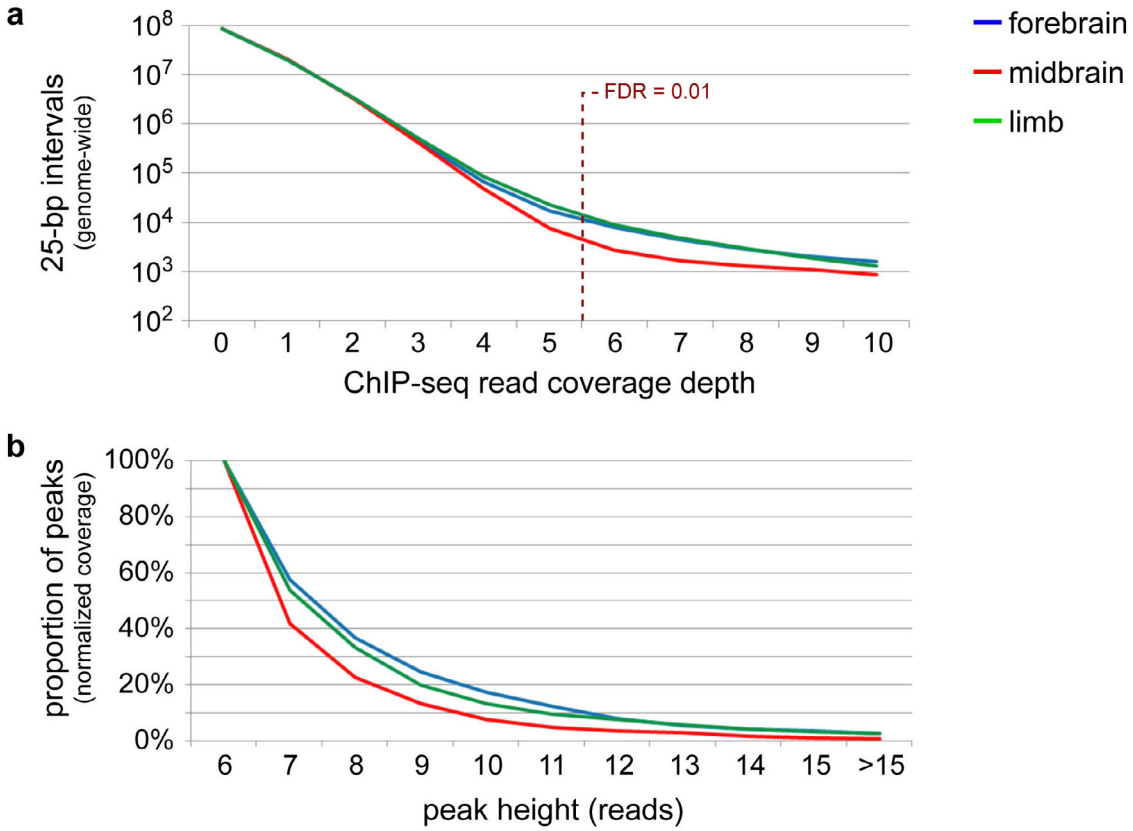
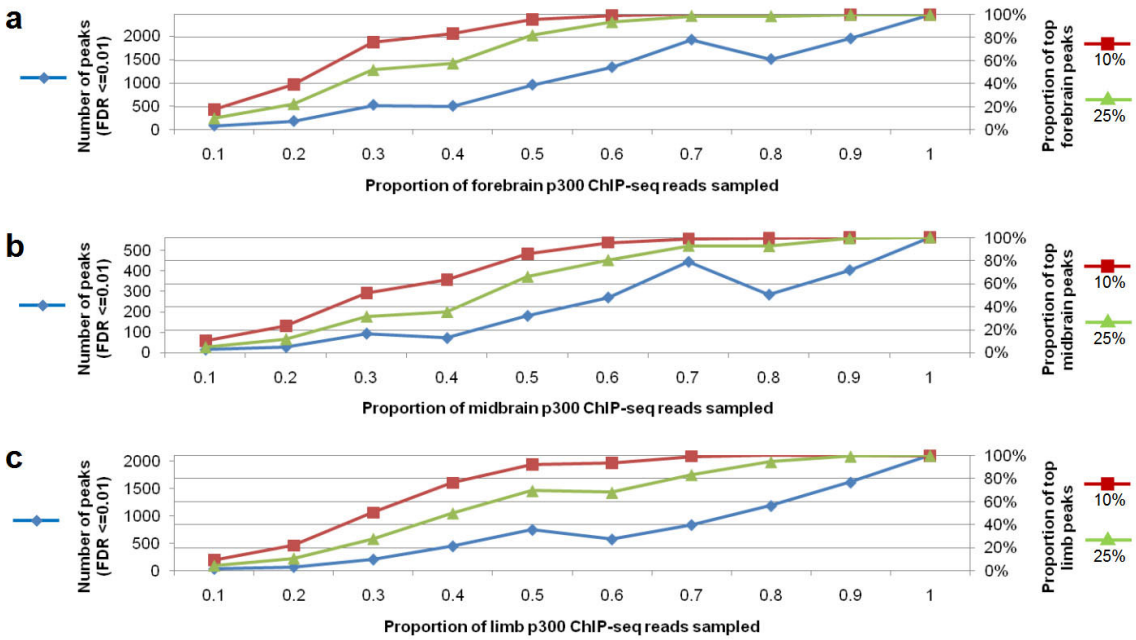
SUPPLEMENTARY INFORMATION



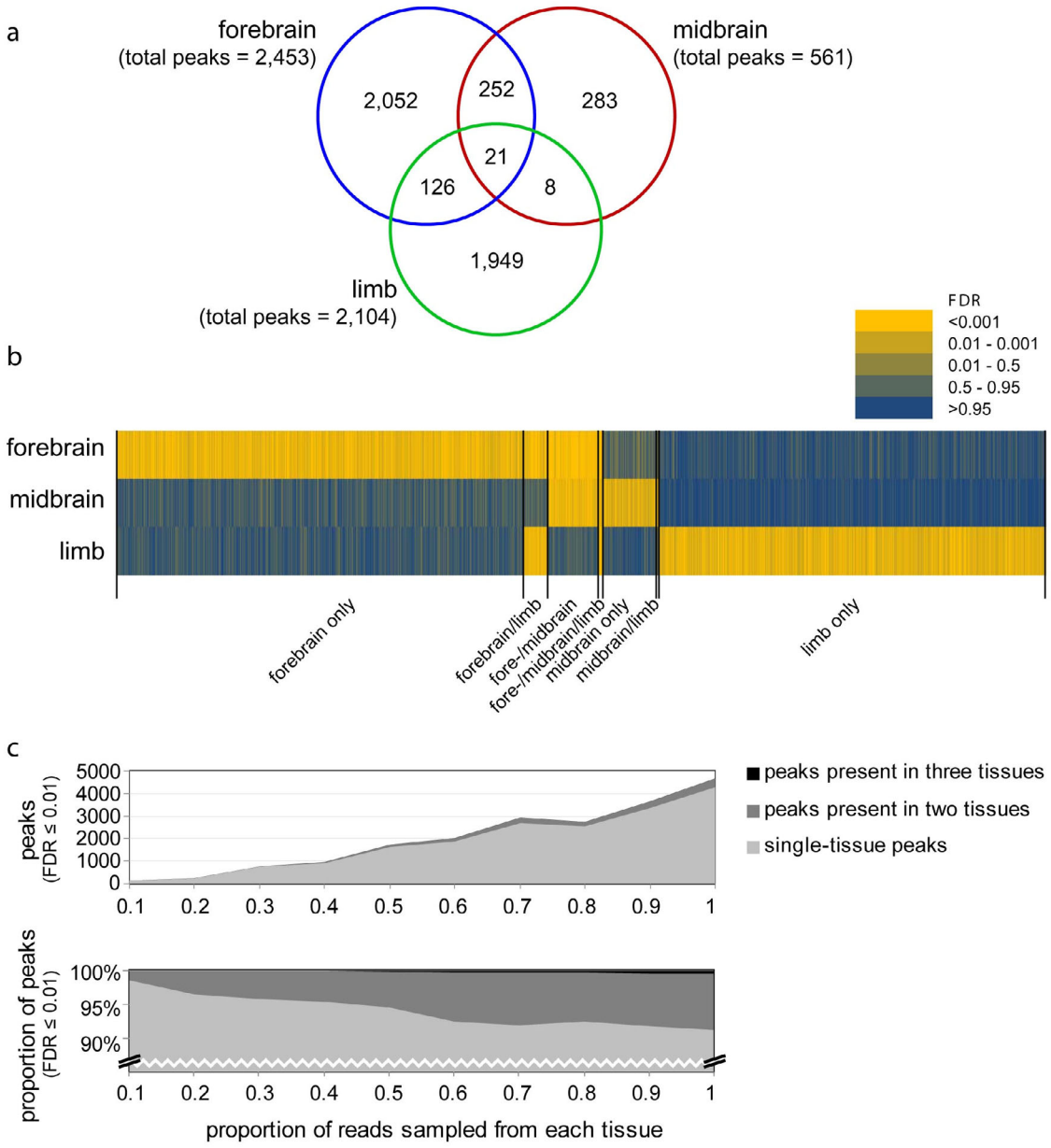
Suppl. Fig. 1: Genome-wide distribution of p300 peaks relative to annotated genes. a) Peaks were classified according to their position relative to UCSC Known Genes. Upstream regions were defined as 20kb upstream of an annotated transcript start site, downstream was defined as 20kb downstream of an annotated transcript end site. Intergenic, intronic and downstream regions were tested in the transgenic mouse assay at an approximately representative ratio, whereas exonic and upstream regions were excluded from transgenic testing. **b)** The distance from the midpoint of each p300 peak to the respectively nearest UCSC Known Genes annotated transcript start site was determined. Peaks within 1kb of the nearest known transcript start site were considered potential promoters and excluded from *in vivo* testing. One tested non-coding element (#1331) is located in an intron of the *Pou2f1* gene, but is here classified as promoter-proximal due to its proximity to an alternative internal transcription start site. Compared to the genome-wide distribution of p300 peaks, the *in vivo* tested elements are mildly skewed towards putative medium-distance (10-100kb) and long-distance (>100kb) enhancers.



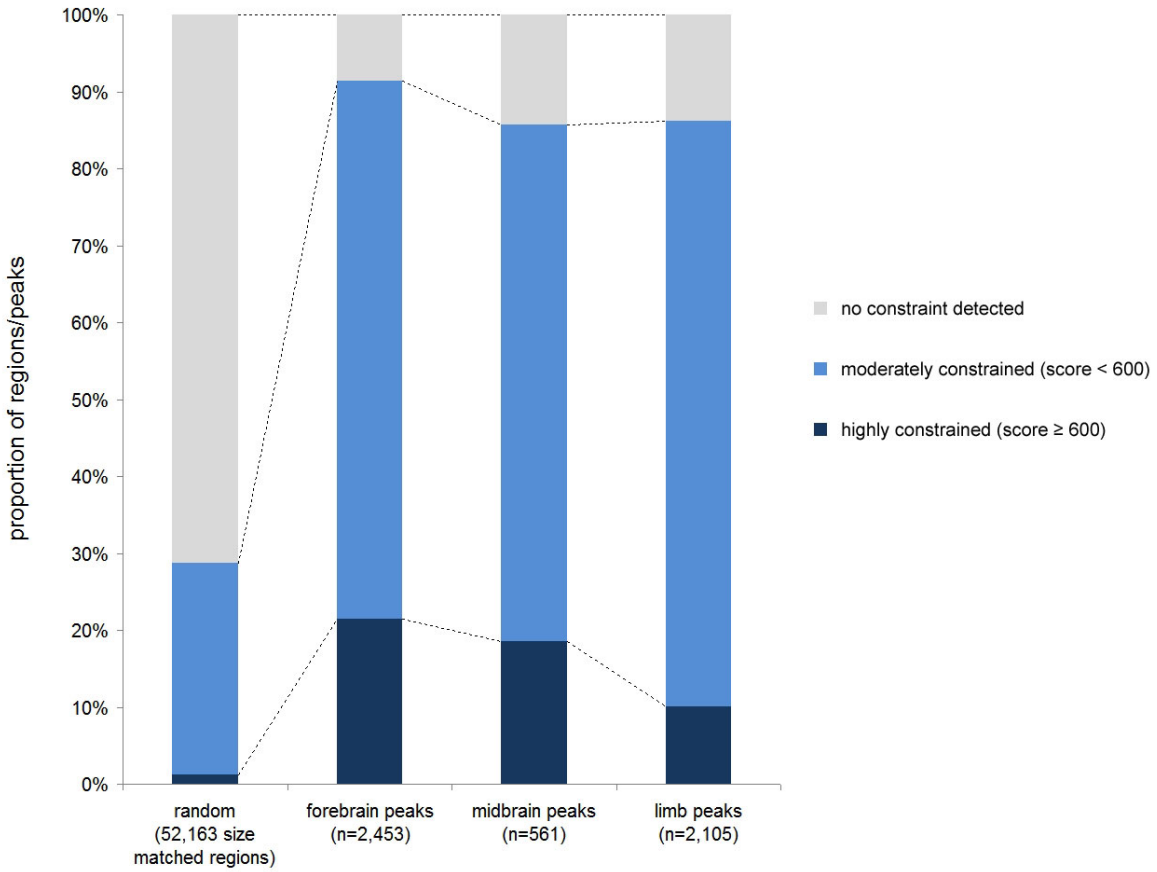
Suppl. Fig. 2: Read clustering and peak coverage properties of p300 ChIP-seq reads from forebrain, midbrain and limb. a) A randomly sampled equal number of reads (n=2,419,480) from each dataset was mapped to the mouse genome, followed by calculation of read coverage in 25bp intervals throughout the genome. The number of peaks with an FDR ≤ 0.01 (i.e., six or more overlapping reads) identified from these identically sized subsets of reads is smaller for midbrain (408) than for forebrain (1,731) and limb (2,105). A smaller fraction of midbrain-derived reads overlap each other compared to forebrain and limb, suggesting less efficient ChIP enrichment prior to massively-parallel sequencing. For example, 7,500 of the 25bp-intervals in the genome are covered by five reads from p300-enriched DNA from midbrain, compared to 17,000 and 22,000 sites in the forebrain and limb datasets, respectively. **b)** The read coverage (peak height) for all significantly enriched regions in each of the three coverage-normalized datasets was determined (cumulative plot). Midbrain-derived p300 peaks are on average covered by fewer reads than forebrain- and limb-derived p300 peaks (average forebrain 7.8 reads, midbrain 7.0 reads, limb 7.6 reads). Consistent with this observation and the overall lower number of peaks from the midbrain sample, the proportion of reads overlapping p300 peaks in the coverage-normalized datasets is lower for midbrain (0.14%) than for forebrain (0.83%) and limb (0.68%).



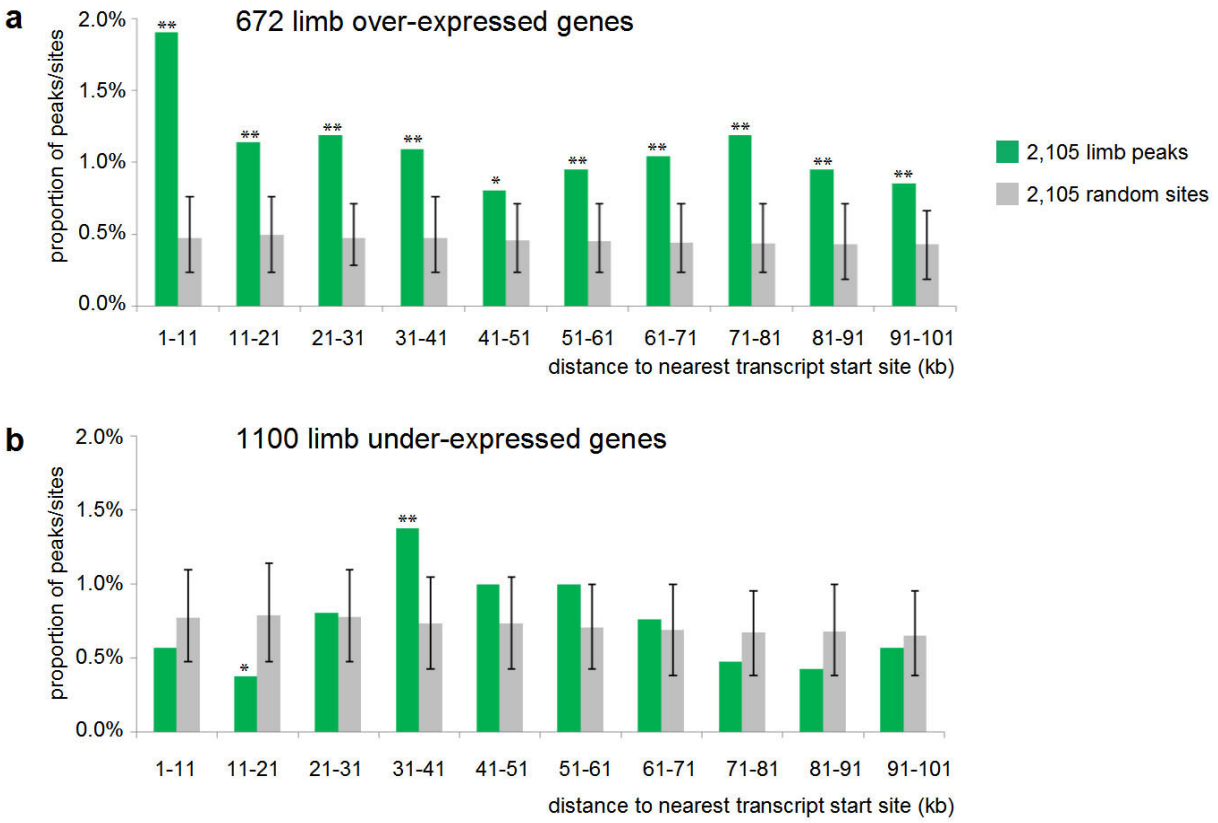
Suppl. Fig. 3: Sensitivity of p300 peak discovery by ChIP-seq increases with sampling depth. For each p300 ChIP-seq dataset, between 10% and 90% of reads were sampled at random and used to calculate peaks at a threshold of FDR ≤ 0.01 (blue, left axis). The number of peaks identified at this threshold does not monotonically increase with increasing sampling depth due to discrete changes in the number or overlapping reads that define FDR = 0.01. To estimate if saturation was obtained for the most significant peaks in each tissue, the resultant peaks for each sample were intersected with the top 10% (red) and top 25% (green) most highly covered peaks from the complete dataset of the respective tissues. In all three tissues, 86%-96% of the eventually most read-enriched peaks (top 10%) were discovered if only half of the data was sampled, suggesting that additional peaks discovered with increasing sampling depth are overall less significantly p300 ChIP-enriched. Each data point represents the average value from 5 individual random samples.



Suppl. Fig. 4: Most p300-enriched regions are tissue-specific. a) Less than 10% of the regions are significantly (FDR<0.01) p300-enriched in more than one of the three tissues, while over 90% are significantly enriched in only one of the three tissues. Most overlap was observed between forebrain and the anatomically adjacent midbrain. **b)** Heat-map representation of all genome regions that are significantly (FDR<0.01) p300-enriched in at least one of the three tissues, clustered into the 7 categories as defined in a). **c)** Re-sampling of subsets of the data indicates that with increasing sequencing depth, the total number of peaks present in only one of the three tissues overall increases (top) while there is a moderate increase in the proportion of regions with significant p300 ChIP-seq read coverage in two or all three tissues examined (bottom).



Suppl. Fig. 5: Most *in vivo* p300-binding regions are significantly constrained in vertebrates. Genome-wide sets of p300-enriched regions in e11.5 forebrain, midbrain and limb were intersected with vertebrate constrained elements [34] and the score (transformed log-odds) of the most constrained element overlapping each of the peaks was considered. For comparison, results for a random set of genome regions (size-matched to forebrain peaks) are shown.



Suppl. Fig. 6: p300 limb peaks are enriched near genes expressed in the limb. We compared the genome-wide distribution of p300-enriched regions in limb tissue at e11.5 with microarray expression data from limb at the same stage [49]. 672 genes were limb-specifically over-expressed and 1100 genes were under-expressed relative to whole embryo RNA at the selected thresholds. Analysis was performed as described for forebrain p300 and microarray data (see Fig. 5 and Methods). a) 10kb bins up to 101kb away from limb over-expressed genes were significantly enriched in limb p300 peaks. b) Consistent with forebrain data, no overall peak enrichment or depletion was observed near limb under-expressed genes except for the 11-21kb bin (depletion) and the 31-41kb bin (enrichment). Note that higher background variation and weaker enrichment in this data compared to forebrain (Fig. 5) may result from differences in the regions of the limb sampled for the gene expression dataset [49] compared with those used for p300 ChIP-seq (Fig. 1). Error bars indicate the 90% confidence interval based on 1000 iterations of randomized distribution (*, $P < 0.05$; **, $P < 0.01$; both one-tailed).

Tissue	Total Reads	Alignable to Mouse	Unambiguously Aligned	Unique Unambiguously Aligned	Peak Coverage Threshold (FDR<0.01)	Peaks Genome-Wide (FDR<0.01)
forebrain	26,759,420	13,728,898 (51%)	8,659,420	3,629,292	7 reads	2,453
midbrain	24,340,547	14,517,733 (60%)	9,431,268	3,530,316	7 reads	561
limb	11,888,250	6,426,526 (54%)	3,950,854	2,419,480	6 reads	2,105
input DNA	39,965,419	13,481,520 (34%)	8,390,111	5,621,346	9 reads	N / A

Suppl. Table 1: Summary of ChIP-seq and mapping results.

			Observed (*in vivo* activity)								
		total transgenic experiments	forebrain	midbrain	limb	forebrain + midbrain	forebrain + limb	midbrain + limb	forebrain + midbrain + limb	positive in other only	negative
Predicted (p300 peaks)	forebrain	*31*	**18**	0	0	6	0	0	0	3	4
	midbrain	*4*	0	**3**	0	1	0	0	0	0	0
	limb	*19*	0	0	**14**	0	0	1	0	1	3
	forebrain + midbrain	*26*	0	0	0	**19**	0	0	2	1	4
	forebrain + limb	*2*	0	0	2	0	**0**	0	0	0	0
	midbrain + limb	*0*	0	0	0	0	0	**0**	0	0	0
	forebrain + midbrain + limb	*4*	1	0	0	0	0	0	**3**	0	0
	Total	***86***									

Correct predictions (Observed pattern exactly matches predicted pattern)	57	66.3%
Partial Predictions (Observed pattern matches prediction in at least one tissue, but missing or unpredicted patterns observed in other tissues)	13	15.1%
Full+partial predictions (combination of the above two categories)	70	81.4%
False positives (At least one predicted pattern was not observed)	19	22.1%
False negatives (At least one observed pattern was not predicted)	10	11.6%

Suppl. Table 6: Predicted and observed *in vivo* enhancer activities.

			feature			
			mappable portion of mouse genome (mm9)	**coding exons***	**introns**	**conserved non-coding sequences****
genome-wide			2,107,016,717bp	33,482,525bp	826,871,215bp	7,926,536bp
proportion of mappable genome			100%	1.6%	39.2%	0.38%
p300 ChIP-seq dataset	**forebrain**	total read bases	1,088,781,607bp	15,305,984bp	378,100,278bp	8,005,450bp
		proportion of read bases	100.0%	1.4%	34.7%	0.74%
		enrichment+	**1.0**	**0.9**	**0.9**	**2.0**
	midbrain	total read bases	1,059,090,633bp	15,802,623bp	368,194,852bp	5,302,923bp
		proportion of read bases	100.0%	1.5%	34.8%	0.5%
		enrichment+	**1.0**	**0.9**	**0.9**	**1.3#**
	limb	total read bases	725,839,855bp	13,215,833bp	260,851,034bp	4,563,299bp
		proportion of read bases	100.0%	1.8%	35.9%	0.63%
		enrichment+	**1.0**	**1.1**	**0.9**	**1.7**

Suppl. Table 7: Genome-wide distribution of ChIP-seq reads. *Coding exons are all complete exons or portions of exons that are translated. ** see methods for definition of conserved non-coding sequences. +Enrichment is calculated as [(fraction of ChIP-seq bases overlapping feature / bases of feature in genome) x bases in mappable genome]. The mappable genome length is estimated as 77.3% of the length of the mouse reference genome sequence (mm9, see methods). # Note that the relatively low enrichment of midbrain p300 reads at conserved non-coding sequences is consistent with fewer peaks and lower ChIP enrichment in this sample.